MCP361 Industrial Engineering Lab: Assignment 6

Due date: 9:00 AM September 4, 2024

- **Naming convention for files for this assignment is as follows**
  MCP361_Entry#_Assignment6_Problem1.py
  MCP361_Entry#_Assignment6_Problem2.py
  MCP361_Entry#_Assignment6.pdf
- **Submit a zip file to Moodle named as follows**
  MCP361_Entry#_Assignment6.zip

Remember the general guidelines for the assignments given at the start of the course.

1. Multi-arm bandit (MAB) problems arise frequently in OR. Your task is to randomly generate a 10-armed bandit problem instance of your choice using python and then try to learn the reward distributions of the ten arms over the time horizon allowed, and you must make use of the greedy action selection algorithm, which is designed to help identify the best arm to pull. You can refer to Chapter 2 of the book "Reinforcement Learning: An Introduction" by Andrew Barto and Richard S. Sutton.

   [4 marks] For this problem, restrict yourself to $T = 500$ time-steps. To collect data that helps you report more representative statistics, it may not be a good idea to run only a single random simulation. Due to this reason, collect data by running $N = 1000$ different problem instances up to $T = 500$ time-steps. Each problem instance must be randomly generated. So, your task is to code the greedy algorithm and run $N = 1000$ simulations (each simulation must run for a period of $T = 500$ time-steps) to collect data and report the statistics in the form of two plots (i) variation of expected reward with experience (ii) fraction of times optimal actions were taken over the learning horizon.

   [1 mark] **Write** down the pseudocode of the algorithm you coded above and **explain** it.

2. [3 marks] Apart from greedy action selection methods, we can use $\varepsilon-$greedy methods that improve learning performance for the 10-armed bandit problem. Code the $\varepsilon-$greedy algorithm in python such that it works for general values of ε. Compare results for the five cases $\varepsilon = 0, 0.01, 0.05, 0.1, 0.25$ using the 10-armed testbed by plotting (i) variation of expected reward with experience (ii) fraction of times optimal actions were taken over the learning horizon.

[2 marks] **Explain** how you chose values of N, T so that you can observe meaningful comparative plots. Also **write** down the insights you gained from these plots.